A brief overview of causal discovery

Jacqueline Maasch

Department of Computer Science

maasch@cs.cornell.edu

7 September 2023

Table of Contents

Causal discovery: Inferring causal structure from data

1 Background

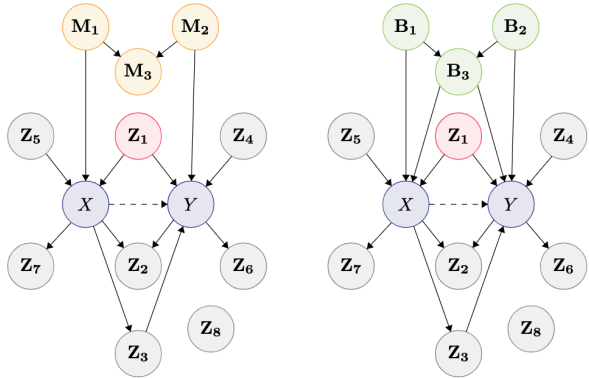


Figure A.2. M-structures and butterfly structures. Ten-node DAG plus M-structure (right) and ten-node DAG plus butterfly structure (left).

The causal structure of a system describes the relations among variables.

Causal discovery: Inferring causal structure from data

- Structure must be (partially) known for causal effect estimation and other forms of inference.

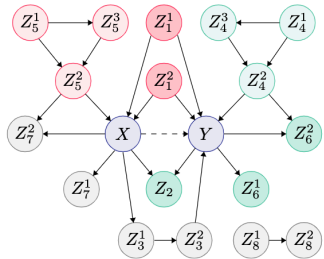
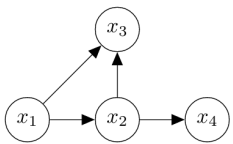


Figure 4: Ten-node DAG with tertiary structures in the form of arbitrary paths between nodes.

- Causal discovery methods infer global and local structures from observational data using statistical and machine learning methods.

Causal relations can be expressed as structural equation models (SEMs) and visualized as directed acyclic graphs (DAGs).

A causal DAG, syn. causal Bayesian network, is a directed graphical model whose nodes represent random variables and whose edges indicate causal relations.



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1.3 & 0 & 0 & 0 \\ 0.3 & 2.5 & 0 & 0 \\ 0 & 5.7 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Figure 1: A Structural Equation Model (SEM). **Left:** DAG with 4 nodes. **Right:** Linear-Gaussian SEM (with $\epsilon \sim \mathcal{N}(0, \Sigma)$)

Causal discovery approaches

1. **Constraint-based methods using conditional independence tests.**
   - PC algorithm. [1]
   - Fast Causal Inference (FCI). [1]
2. **Score-based methods using optimization.**
   - Greedy Equivalence Search (GES). [2]
   - DAGs with NO TEARS. [3]
3. **Functional causal models.**
   - Nonlinear additive noise models (ANM). [4]
   - Post-nonlinear additive noise models (PNL). [5]
   - Linear non-Gaussian acyclic models (LiNGAM). [6]
4. **Other methods that exploit asymmetries in the data.**
   - Information-geometric causal inference (IGCI). [7]

Table of Contents

2 Constraint-based methods

Constraint-based causal discovery

2 Constraint-based methods

Premise:

Assumes equivalence between properties of the data and properties of the causal graph:

- **Causal Markov:** d-separation in graph $\mathcal{G}$ implies conditional independence in data distribution P.

$$X \perp\!\!\!\perp_{\mathcal{G}} Y|Z \Rightarrow X \perp\!\!\!\perp_{P} Y|Z$$

- **Faithfulness:** conditional independence implies d-separation.

$$X \perp\!\!\!\perp_{P} Y|Z \Rightarrow X \perp\!\!\!\perp_{\mathcal{G}} Y|Z$$

Pros and cons

Pros:

- Well-established theory.

Cons:

- Faithfulness is a strong assumption.
- Independence testing is a challenge in itself, and may require large sample sizes.
- Total tests exponential in node count in worst case.
- Cannot handle two-variable case.
- Identifies up to a Markov equivalence class.

PC algorithm

2 Constraint-based methods

Premise

Assumptions: Causal Markov, faithfulness, causal sufficiency [8].

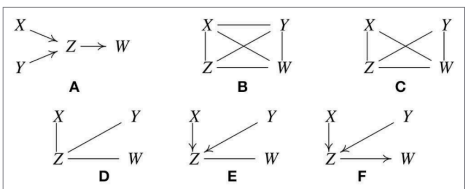


FIGURE 1 | Illustration of how the PC algorithm works. **(A)** Original true causal graph. **(B)** PC starts with a fully-connected undirected graph. **(C)** The $X - Y$ edge is removed because $X \perp\!\!\!\perp Y$. **(D)** The $X - W$ and $Y - W$ edges are removed because $X \perp\!\!\!\perp W \mid Z$ and $Y \perp\!\!\!\perp W \mid Z$. **(E)** After finding v-structures. **(F)** After orientation propagation.

Pros

Pros:

- Straightforward.
- No innate parametric assumptions, though independence tests (usually) make some.

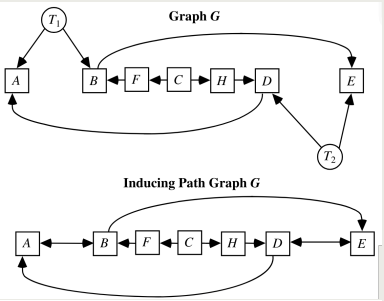
Cons:

- Returns a Markov equivalence class instead of a unique DAG.
- Bottlenecked by challenges in conditional independence testing.
- Cannot handle latent confounding.

Fast Causal Inference (FCI)

2 Constraint-based methods

Ground truth

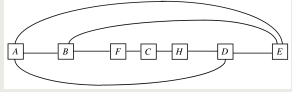


Definition 5 (inducing path) *A path between two nodes X and Y is an inducing path if and only if every internal node in the path is a collider on the path and is an ancestor of X or Y or both.*

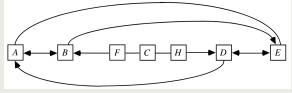
FCI [1]

A PC extension that is asymptotically correct in the presence of latent confounding.

1: Remove edges from full undirected graph (like PC).



2: Orient edges by ID'ing colliders (like PC).



3: Unorient then partially reorient edges (not like PC).

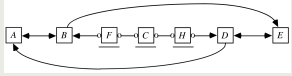


Table of Contents

3 Score-based methods

Score-based causal discovery

3 Score-based methods

Premise

Search for DAG that best fits data using **scoring function** $\mathcal{S}$, s.t. $\hat{\mathcal{G}} := \text{argmax}_{\mathcal{G}}\, \mathcal{S}(\mathcal{D}, \mathcal{G})$.

- Traditionally, this is a **combinatorial optimization** problem, but can be converted to **continuous optimization**.
- **Common scoring functions**: Bayesian Information Criterion (BIC), Minimum Description Length (MDL; approximates Kolmogorov Complexity), Bayesian Gaussian equivalent (BGe), Bayesian Dirichlet equivalence (BDe), etc.

Pros and cons

Pros:

- No longer reliant on independence testing.
- For continuous optimization, can employ modern black-box optimizers.

Cons:

- Challenges from combinatorial opt. and asymmetric directed adjacency matrices.
- Must choose a good scoring function.
- Theoretical guarantees for finite data?
- Search space grows superexponentially with node count, so it must be restricted.

Greedy Equivalence Search (GES)

3 Score-based methods

Premise

2-phase greedy search over equivalence classes using Bayesian scoring criterion [9]:

1. Start with empty graph and iteratively add edges until goodness-of-fit reaches local maximum.
2. Iteratively eliminate edges until score reaches local maximum.

Pros and cons

Pros:

- Guarantees under infinite data: global minimizer when assumptions are met.

Cons:

- Returns Markov equivalence class.
- No guarantees under finite data.
- Assumes causal sufficiency.

GES: An optimal two-phase algorithm

3 Score-based methods

GES entails two phases:

1. **Forward Equivalence Search (FES).** Starting with the equivalence class of no dependencies, greedily make single-edge additions until a local maximum is reached.
2. **Backward Equivalence Search (BES).** Consider all single-edge deletions within the current equivalence class until a local maximum is reached.

Chickering shows that GES correctly identifies the optimal solution in the large sample limit when applied to the sparsely-connected search space of equivalence classes.

1. **The local maximum reached after FES contains the generative distribution.** Proof follows from the assumption that p is DAG-perfect.
2. **The equivalence class reached after BES must be a perfect map of the generative distribution.** Proof follows from Theorem 4.

Table of Contents

4 Functional causal models

Functional causal models (FCM)

Premise: Exploit asymmetries

- Represents effect Y as a function of direct causes X and unmeasurable noise ϵ, e.g.: $Y = f(X) + \epsilon$.
- Assume X, Y have a direct causal relationship and no confounders + additional parametric assumptions.
- Fit the FCM for both causal directions. Test for independence between estimated noise and cause.
- **The direction which gives finds the hypothetical cause and noise terms to be independent is considered plausible.**
- Can use nonlinear regression of Y on X to obtain $\hat{f}$, residuals $\hat{\epsilon} = Y - \hat{f}(X)$.

Pros and cons

Pros:

- Well suited for continuous case.
- Handles bivariate inference and larger structures.

Cons:

- Linearity assumption of LiNGAM variants [6] limits applicability to real-world data.
- Nonlinear additive noise models [4] tolerate arbitrary causal functional forms, but increased applicability comes with high complexity.

Bivariate causal direction inference

4 Functional causal models

- Conditional independence testing requires at least three random variables for the simplest case, $X \perp\!\!\!\perp Y|Z$, so it cannot identify the bivariate causal model.
- Imposing functional and distributional assumptions on the causal model can reveal statistical and information-geometric asymmetries that enable identifiability [4–7].

	Method	***Description***	***Parametric assumptions***
1.	Post-nonlinear additive noise model (PNL) [46]	Most general form	Arbitrary functional and distributional forms
2.	Nonlinear additive noise model (ANM) [45]	Special case of PNL	Arbitrary functional and distributional forms
3.	Linear Non-Gaussian Acyclic Model (LiNGAM) [47]	Special case of PNL	Linear mechanisms and non-Gaussian errors
4.	Information-geometric causal discovery (IGCD) [44]	Does not assume PNL	Arbitrary functional and distributional forms

Table 4: Methods for bivariate causal direction inference.

Post-nonlinear additive noise model (PNL)

The post-nonlinear model (PNL) is the most general of the well-defined functional causal models. It is suited for bivariate inference and inference on larger structures.

The PNL accounts for the nonlinear effect of the cause, the inner noise effect, and the measurement distortion effect in the observed variables.

Under this model, each variable x_i in graph $\mathcal{G}$ takes the form

$$x_i = f_{i2}\left[f_{i1}(pa_i) + \epsilon_i\right] \tag{1}$$

where pa_i are the parents of x_i, ϵ_i is the noise term, and f_{i1}, f_{i2} are arbitrary functions.

Post-nonlinear additive noise model (PNL)

The PNL comes with thorough identifiability results. Most notably, it is **not identifiable under the linear Gaussian setting.**

The identifiability conditions of the PNL extend to its special cases, which include:

- **Linear additive models (e.g. LiNGAM):** Given $x_i = f_{i2}\left[f_{i1}(pa_i) + \epsilon_i\right]$, f_{i1} is linear and f_{i2} is the identity function.
- **Nonlinear additive models (e.g. ANM):** f_{i1} is nonlinear and f_{i2} is the identity function.
- **Multiplicative noise models:** The multiplicative noise model $x_i = pa_i \cdot \epsilon_i$ can be expressed as $\exp(\log pa_i + \log \epsilon_i)$ where $f_{i1}(pa_i) = \log(pa_i)$ and $f_{i2}(\cdot) = \exp(\cdot)$.

Table of Contents

5 Local causal discovery

Thinking locally

5 Local causal discovery

Global causal discovery is hard. Local causal discovery can simplify things.

- What if we don't need to know the entire causal structure?
- Can we increase efficiency or improve inference by zooming in only on the substructures we need?

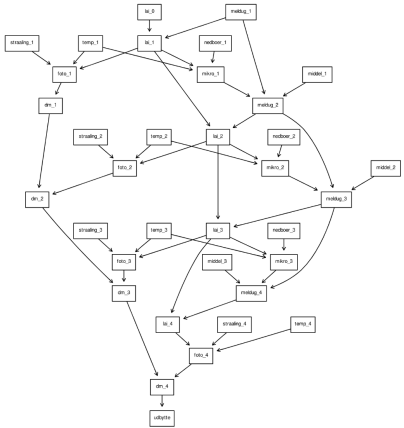


Figure A.8: The complete ground truth MILDEW DAG obtained from `bnlearn`. The subgraph used for inference in our benchmarking experiments in visualized in Figure A.6.

Thinking locally

5 Local causal discovery

Global causal discovery is hard. Local causal discovery can simplify things.

- What if we don't need to know the entire causal structure?
- Can we increase efficiency or improve inference by zooming in only on the substructures we need?

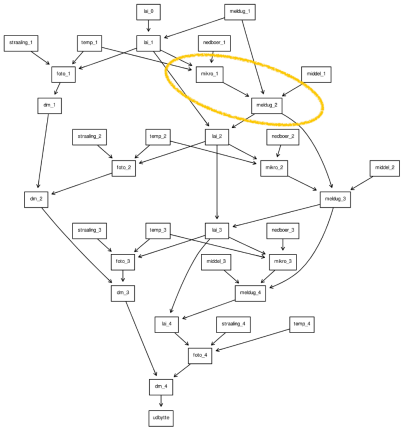


Figure A.8: The complete ground truth MILDEW DAG obtained from `bnlearn`. The subgraph used for inference in our benchmarking experiments in visualized in Figure A.6.

Local causal discovery around exposure-outcome pairs

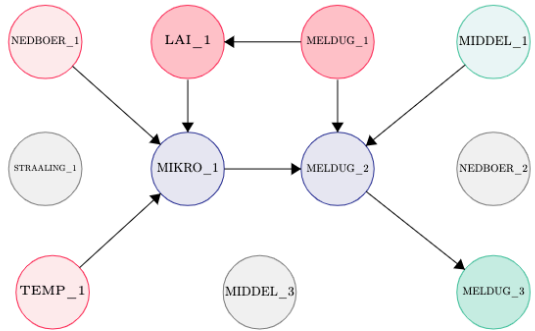


Figure A.6: Causal structure of the MILDEW benchmark. The full ground truth DAG is pictured in Figure A.8.

Thank you! Any questions?

maasch@cs.cornell.edu

Table of Contents

References

6 References

[1] P. Spirtes et al. *Causation, Prediction, and Search*. en. Ed. by J. Berger et al. Vol. 81. Lecture Notes in Statistics. New York, NY: Springer New York, 1993. DOI: 10.1007/978-1-4612-2748-9.

[2] D. M. Chickering. "Statistically Efficient Greedy Equivalence Search". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*. 2020.

[3] X. Zheng et al. "DAGs with NO TEARS: Continuous Optimization for Structure Learning". en. In: (), p. 12.

[4] P. O. Hoyer et al. "Nonlinear causal discovery with additive noise models". en. In: *Advances in Neural Information Processing Systems 21 (NIPS 2008)*. 2008.

[5] K. Zhang et al. "On the Identifiability of the Post-Nonlinear Causal Model". en. In: *Uncertainty in Artificial Intelligence* (2009).

References

6 References

[6] S. Shimizu et al. “A Linear Non-Gaussian Acyclic Model for Causal Discovery”. en. In: *Journal of Machine Learning Research* 7 (2006), pp. 2003–2030.

[7] D. Janzing et al. “Information-geometric approach to inferring causal directions”. en. In: *Artificial Intelligence* 182-183 (2012), pp. 1–31. DOI: 10.1016/j.artint.2012.01.002.

[8] C. Glymour et al. “Review of Causal Discovery Methods Based on Graphical Models”. en. In: *Frontiers in Genetics* 10 (2019), p. 524. DOI: 10.3389/fgene.2019.00524.

[9] D. M. Chickering. “Optimal Structure Identification With Greedy Search”. en. In: *Journal of Machine Learning Research* 3 (2002).